ICIP 2002 Proceedings

COMPRESSION OF cDNA AND INKJET MICROARRAY IMAGES

*Rebecka Jornsten, Bin Yu**

University of California, Berkeley
rebecka@, binyu@stat.berkeley.edu

Wei Wang, Kannan Ramchandran†

University of California, Berkeley
wangwei@, kannanr@eecs.berkeley.edu

ABSTRACT

Microarray image technology is a powerful tool for monitoring the expression of thousands of genes simultaneously. Each microarray experiment produces immense amounts of image data, and efficient storage and transmission requires compression that utilizes microarray image's structure and unique analysis goals. Hence, we have developed a progressive compression scheme for microarray images which can be either lossy or lossless. Our scheme has a coded data structure that allows fast decoding and reprocessing of image *subsets*, and includes summary statistics and image segmentation information. Since visual fidelity is *not* the end goal for microarray images, we introduce a new measure of distortion for lossy compression: the sensitivity of microarray information extraction to compression loss. We find that a lossy compression ratio of 8:1 for cDNA microarrays minimally affects downstream processing. The average lossless compression ratio is 1.83:1 for cDNA images and 2.43:1 for inkjet images, comparable to state-of-the-art lossless schemas, yet with added flexibility and information.

1. INTRODUCTION

Microarrays have become an important tool for developing understanding of gene function, regulation and interaction through the simultaneous study of thousands of genes. The raw data produced by a microarray experiment is an image (or more accurately, a two-dimensional matrix of intensity values) of a grid of thousands of spots, one for each gene. [1] The proven and potential scientific value of microarray image technology is enormous, but it is widely acknowledged that the quality of microarray data is highly variable. Without standardization of methods and microarray processing tools, results from different labs are rarely comparable. We need to distinguish between the inherent noise of the data, the source of which lies in experimental variation, and the noise introduced by different methods used for genetic information extraction. For this reason, it is often advisable to store the raw image data from an experiment rather than just the summary statistics from a particular extraction algorithm. Moreover, microarray experiments are costly and the image data files are large (>40 MB). Various organizations are establishing publicly accessible databases for sharing microarray data. The cost of data storage is rapidly decreasing. However, for efficient transmission and data distribution, and for storage of large quantities of image data, compression is an essential tool.

We propose a progressive compression scheme which can be lossless or lossy. We add necessary flexibility to traditional lossless coding that facilitates downstream genetic information extraction analysis. In lossy compression, we preserve the genetic information in the *lossily compressed* microarray images, such that the downstream tasks are unaffected by the compression. We take into consideration that the appropriate measure of performance for compression of microarray images is not visual fidelity or MSE, as is generally the case in image compression. Microarray images are processed using a multi-step procedure (image segmentation, information extraction, normalization), and thus we do not have a simple distortion criterion over which to optimize. One variant of our lossy scheme has a locally varying bound on the maximum pixel-wise error, i.e. the error bound is a function of the local signal-to-noise ratio. With this approach, at low bit-rates (compression ratio 8:1 for cDNA microarrays), the effect of lossy compression on the extracted genetic information is smaller than the array-to-array variability in replicated experiments. Moreover, the impact of compression is smaller than that of changing the method of information extraction. Thus we define a new criteria of acceptable distortion for lossy compression of microarray images.

Microarray imaging is an emerging technology and several experimental procedures have been developed producing different image characterizations, including cDNA, oligonucleotide, and inkjet technologies. We will describe compression results on cDNA and inkjet microarrays.

*NSF FD98-02314, NSF DMS-9803063, ARO DAAG55-98-1-0341, NSF FD01-12731

†NSF graduate fellowship, Lucent Grant, NSF Career Award

[1] Section 2 discusses the mechanics of microarray experiments.

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

20021022 066

2. cDNA AND INKJET MICROARRAY IMAGES

Messenger RNA (mRNA) acts as an intermediate in protein synthesis from genes contained in DNA. The amount of mRNA present in a cell is related to the gene expression level. Microarray experiments measure differential gene expression, or the amount of genetic material in a sample *relative* to another sample, through competitive hybridization. In the cDNA microarray procedure, DNA probes (each corresponding to a gene) are "spotted" onto a glass slide by a robotic *arrayer*. Then mRNA samples from two different cells are labeled with fluorescent tags (commonly referred to as "red" and "green"), and are mixed and hybridized onto the array. A laser scan of the array produces two fluorescent intensity images. The intensity ratio for each probe, or spot, is proportional to the relative abundance of hybridized mRNA in the two samples.

Inkjet microarrays employ the same concept but uses synthesized oligonucleotide probes and printer technology to deposit the material onto a slide.

The images are structured, with high intensity spots (corresponding to the probes) located on a grid (see figure 1). The spots are submerged in a noisy and non-stationary background. The spots have roughly circular shape, though some show significant deviation from this shape due to the experimental variation of the spotting procedure.

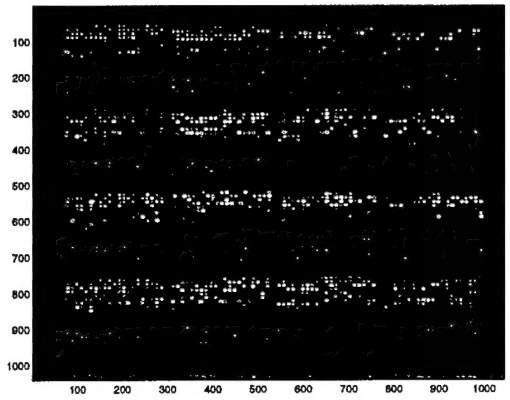

Fig. 1. Microarray image: 4×4 grid set-up, 19×21 spots per grid.

2.1. Genetic information extraction

To estimate the differential gene expression, we have to identify the high intensity regions in the image corresponding to each probe (image segmentation), and estimate and remove the local background noise (background correction).

Automatic registration of the image is used to determine the approximate centers, or the grid location, of the spots. [1][2]. We use a seeded region growing algorithm for initial segmentation, followed by a two-component Gaussian mixture model, to further refine the boundaries of the spots, or regions of interest (ROI) [3]. Various approaches to background correction are described in ([2], [4], [1]).

An example of spot, or gene, summary statistics for cDNA is the ratio of background corrected mean intensities. We denote by R_i the red (fluor tag) scan pixels, and by G_i the green scan pixels. The differential expression level, R/G, is then calculated as the ratio of the mean ROI intensities:

$$\frac{R}{G} = \frac{\frac{1}{S}\sum_{R_i \in ROI} R_i - BgR}{\frac{1}{S}\sum_{G_i \in ROI} G_i - BgG},$$

where Bg refers to the estimates of the local background, and S is the number of ROI pixels.

Normalization is necessary prior to statistical modeling and array-to-array analysis. [5]

3. LOSSY AND LOSSLESS COMPRESSION OF MICROARRAY IMAGES

Background correction and normalization involve segmenting foreground from background, and microarray backgrounds are invariably low-intensity. To facilitate this segmentation, a compression scheme must keep more precision for low intensity spot regions, but can use a coarse image reconstruction for high intensity spot regions. Note that the commonly used MSE criterion does not reflect this. We find that local SNR is a good indicator for background-corrected spot intensity. Thus we vary the pixel-wise error bound using local SNR thresholds. Our aim is to keep both bias and variance under control, and ensure that the effect of compression is smaller than the variability between replicated experiments, or between methods of information extraction to another. We define this as *acceptable loss* for microarray image compression.

Ideally we want the output of the compression scheme to have a data structure. This would enable us to transmit, and reconstruct image subsets at different precision.

State-of-the-art lossy compression schemes for natural and medical images are almost without exception wavelet based. However, this is not appropriate for the compression of microarray images. In fact, the many and small high intensity regions create large wavelet coefficients over almost the entire image subbands. At low bit-rates, algorithms such as SPIHT, or wavelet and zero-tree coding [6] will be dominated by the edges around the high intensity spots. Compression schemes that are not wavelet based, but based on predictions in the spatial domain, also have difficulty with the many high intensity spots. A rowscan based prediction scheme creates a "smearing" bias in the image reconstruction.

We take a spatial prediction approach. To avoid the "smearing" bias we encode the spot regions and the background separately. We first transmit an *overhead* defining the ROI and background, i.e. a segmentation map. We refer to this approach as segmented LOCO, or SLOCO.

3.1. Segmented LOCO (SLOCO)

Our scheme builds on the JPEG 2000 lossless standard, LOCO (LOw COmplexity), [7]. Here we will assume the reader is familiar with the components of LOCO and will briefly discuss how our method differs.

SLOCO differs from standard LOCO in mainly three aspects. Firstly, the spots and backgrounds are encoded separately. Secondly, a UQ-adjust quantizer is used instead of the UQ quantizer. Thirdly, we allow for varying maximum pixel-wise error bounds δ. We use a runlength code that takes the segmentation map, and the varying δ into account.

We use a UQ-adjust quantizer for large δ. Thus, the bins near the center of the error distribution, where most of the probability mass is located, have adjusted reconstruction levels closer to the MSE distortion optimal, as well as smaller bin-widths $2\delta' + 1$, such that the maximum error is still bounded by δ. The outer bins have bin-widths $2\delta + 1$, and center bin reconstruction levels.

3.2. Overhead

The overhead of the SLOCO algorithm contains the spot means and standard deviations, as well as the local background intensity estimates and estimates of the local background standard deviation. The overhead also contains the estimated grid structure of the microarray images, and the segmentation map.

The spot and background means are encoded using adaptive Lempel-Ziv. The segmentation map is efficiently encoded using the chain code of Lu and Dunham [8]. The average cost of the overhead is ~ 0.376 bpp for the cDNA data and 0.076 for the inkjet data.

If no re-processing of the images is needed, the overhead contains all relevant information for down-stream analysis. In addition, it contains spot quality measurements such as spot shapes, variances and the local background variance.

3.3. Coding the Spot Regions

Given the overhead we can compute the signal-to-noise ratio of each spot. Based on the SNR we can pick a bound on δ for each spot. For cDNA, the size of each spot is too small to allow for any adaptive prediction step, or for adaptive estimation of the Golomb parameter. We therefore use a fixed Golomb code, and only the fixed predictor $\hat{x}_{fix}$ within each spot. The spot Golomb parameter could be estimated on the encoder side, after applying the fixed predictor, and transmitted as overhead to the decoder. However, we can do nearly as well by using an approximate estimate of the optimal Golomb parameter k. We encode the spots in a row scan manner. Missing context pixels for the predictor are filled in with the spot mean from the overhead.

3.4. Coding the Background

The background is encoded in a row scan fashion for sub-blocks of the images. It is more efficient to encode background pixels for the entire image as one block. However, we find that encoding sub-blocks (in cDNA, corresponding to the 4×4 print-tip configuration) gives approximately the same bit-rate but allows for image subset reconstruction.

Missing context pixels, i.e. the spot pixels, are filled-in using local background intensity estimates.

The runlength coding of SLOCO differs from regular LOCO in that we do not allow runs to cross from a region with higher maximum error bound δ, into one with smaller δ. If a spot is encountered during a run, we skip ahead to the next background pixel.

4. PROGRESSIVE TRANSMISSION

Our lossy compression scheme can be extended to a fully lossless reconstruction of the microarray images. Given the initial lossy reconstruction, the image reconstruction can be refined, spot by spot, background region by background region, or even pixel by pixel, to any bit-rate above the minimum decodable bit-rate. Our scheme is thus progressive.

The overhead provides us with the maximum error δ in each region of the lossy reconstruction of the images. We encode bit planes of the residual image. If the maximum error in a region chosen is δ, then the quantization errors are approximately uniformly distributed on $[-\delta, \delta]$. This holds if δ and δ' are reasonably close, and small. A UQ with two reconstruction levels corresponds to the first bit plane, and equals the sign of the quantization errors, which can be sent at rate 1 bpp. The decoder then reconstructs a refined error as $sign \times \delta/2$ and adds this to the previous lossy reconstruction. If we want to refine this region of the image further, we compute a new residual image, again send the sign of the residual to the decoder, and reconstruct at levels $sign \times \delta/4$. This achieves the $\log_2(2\delta + 1)$ bit-rate. We choose this simple coding scheme since it gives us total freedom to encode any part of the residual image at any rate we desire, independently of what we choose to do in other regions of the image.

If δ and δ' are very dissimilar, the quantization errors will not be uniformly distributed. We then compute the quantization bin centroids at the encoder, and send as overhead to the decoder, with negligible increase in coding cost for the large microarray images.

5. RESULTS AND COMPARISON OF METHODS

The average LOCO lossless compression ratio for our test cDNA images is 1.85:1. The SLOCO ratio is 1.83:1 (including the overhead information). In comparison, Lempel-Ziv (gzip) gives a compression ratio of 1.48:1, SPIHT 1.65:1, and wavelet (zero-tree coding + entropy coding of residual) 1.72:1. The 8 least significant bits of cDNA images are close to random, i.e. have marginal entropy 8 bpp, and are unpredictable. This puts a ceiling of 2:1 on the lossless compression ratio. The average LOCO lossless compression ratio for inkjet images is 2.44:1, while the SLOCO ratio is 2.43:1.

Note that the SLOCO encoding contains easily extractable summary statistics and segmentation information not available from LOCO. Moreover, SLOCO allows fast reconstruction of subsets of the image, allowing researchers to access individual spots out of hundreds of thousands per image.

The lossy compression ratio, using variable pixel-wise distortion bound, is 8:1 for cDNA microarray images. We find that the variability introduced by the lossy compression, in the extraction of the differential gene expression levels, is smaller than the array-to-array variability [9]. We also find that the difference in extracted gene expression levels between different methods of genetic information extraction is much greater than the difference between lossless and lossy reconstructions of the images. For inkjet images, we have not had access to downstream analysis. The lossy compression ratio for inkjet, with fixed maximum pixel-wise error $\delta = 8$ (3 bpp), is 6.51:1 using LOCO and 6.44:1 using SLOCO.

6. CONCLUSION AND FUTURE WORK

We have presented a lossy and progressively lossless compression scheme for microarray images. The flexible structure of our scheme allows for lossless, or refined precision reconstruction for any subset of the images. At a compression ratio of 8:1 for cDNA images, we find that the tasks of genetic information extraction with a variety of methods are only marginally affected by the compression. The effect of compression is smaller than the array-to-array variability. The effect is also smaller than the difference between alternative methods for information extraction. In fact, compression can improve the estimation of gene expression levels for cDNA images. We found that compression acts as a form of shrinkage for large absolute gene expression levels toward the mean of replicated arrays (taken as "truth").

As future work, we plan to formalize our bit allocation scheme in lossy SLOCO into a rigorous information theoretic framework. Under this framework, we require a distortion function that measures the sensitivity of downstream information extraction to compression. Using this sensitivity function, we allocate bits among the different gene spots. To encode m independent random sources $X_1, \ldots, X_m$ (e.g. the spots) with R bits, the classic problem is how to allot these bits to the different components to minimize the total distortion or sensitivity. For example, using local SNR as an indicator of this sensitivity, an analogy between our SNR-thresholded variable pixel-wise error bound and reverse water-filling seems natural. We spend more bits (by lowering the pixel-wise error) on spots with 1/SNR values greater than a threshold

7. ACKNOWLEDGEMENTS

We thank P. Brown (Department of Biochemistry, Stanford University), Lawrence Livermore National Lab, and P. Lun (Department of molecular and cell biology, UC Berkeley) for providing us with cDNA data, and E. Schadt (Rosetta Inpharmatics) for providing us with inkjet data.

8. REFERENCES

[1] Y. Yang, M.J. Buckley, S. Dudoit, and T. Speed, "Comparisons of methods for image analysis on cdna microarray data," Tech. Rep., Statistics, UC Berkeley, Berkeley, California, 2000.

[2] M. Eisen, "Scanalyze," `http://rana.stanford.edu/software`, 1998.

[3] R. Adams and L. Bischof, "Seeded region growing," *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 16, pp. 641–647, 1994.

[4] Genepix-Axon Instruments, "Genepix 4000a, user's guide," 1999.

[5] Y. Yang, S. Dudoit, P. Luu, and T. Speed, "Normalization for cdna microarray data," Tech. Rep., SPIE BiOS, San Jose, California, Jan 2001.

[6] A. Said and W. Perlman, "Reversible image compression via multiresolution representation and predictive coding," *Proc SPIE*, vol. 2094, pp. 664–674, 1993.

[7] M. Nerhav, G. Seroussi, and M. Weinberger, "Modeling and low-complexity adaptive coding for image prediction residuals," *Int'l Conference on Image Processing, Lausanne*, 1996.

[8] C. Lu and J. Dunham, "Highly efficient coding schemes for contour lines based on chain code representation," *IEEE trans Comm*, vol. 39, no. 10, pp. 1511–1514, 1997.

[9] R. Jornsten, "Data compression and its statistical implications," *Ph.D. Thesis*, 2001.